Analysis of Complexity and Modulation Spectra parameterizations to characterize Voice Roughness

Laureano Moro-Velazquez , Jorge Andrés Gómez García, Juan Ignacio Godino-Llorente,

Abstract. Disordered voices are frequently assessed by speech pathologists using acoustic perceptual evaluations. This might lead to problems due to the subjective nature of the process and due to the influence of external factors which compromise the quality of the assessment. In order to increase the reliability of the evaluations the design of new indicator parameters obtained from voice signal processing is desirable. With that in mind, this paper presents an automatic evaluation system which emulates perceptual assessments of the roughness level in human voice. Two parameterization methods are used: complexity, which has already been used successfully in previous works, and modulation spectra. For the latter, a new group of parameters has been proposed as Low Modulation Ratio (LMR), Contrast (MSW) and Homogeneity (MSH). The tested methodology also employs PCA and LDA to reduce the dimensionality of the feature space, and GMM classifiers for evaluating the ability of the proposed features on distinguishing the different roughness levels. An efficiency of 82% and a Cohen's Kappa Index of 0.73 is obtained using the modulation spectra parameters, while the complexity parameters performed 73% and 0.58 respectively. The obtained results indicate the usefulness of the proposed modulation spectra features for the automatic evaluation of voice roughness which can derive in new parameters to be useful for clinicians.

Keywords: GRBAS, Complexity, Modulation Spectra, Kappa Index, GMM, voice pathology, Roughness.

1 Introduction

Voice pathology assessment aims at diagnosing and evaluating the condition of patients with vocal pathologies, in order to find an appropriate treatment for their disorders. On this context, speech pathologist often employ perceptual analysis of patient's phonation to indicate the perceived level of perturbation of the voice. In these cases, specialists listen to the voice of the patient producing a sustained vowel or reading a particular passage and rate it conforming to a specific procedure. Most of the times, a numeric value is assigned according to the dysfunction level, where one of the most used rating scale is $GRBAS$ [1]. This scale is divided into five traits which evaluate different speech quality characteristics: Grade (G), Roughness (R), Breathiness (B), Aesthenia (A) and Strain (S).

Each characteristic ranges from 0 to 3, where 0 indicates no affection, 1 slightly affected, 2 moderately affected and 3 severe affected voice regarding to the corresponding trait.

The main issue affecting the perceptual analysis of voice quality is the inherent subjectivity of the process, where external factors might compromise the quality of the assessment, such as the assessor's mood, its background training, fatigue, stress or cultural issues, among many others [2, 3].

With this in mind, acoustic analysis of voice signal techniques might be considered for reducing the uncertainty of perceptual evaluations. The acoustic analysis is widely used as a tool for monitoring the patient's evolution after the application of a treatment, mainly due to the simplicity of the process, as well as its low cost and non-invasiveness. Therefore its use in quality assessment of voice might be beneficial to clinicians, giving them tools to perform evaluations in a more objective and reproducible manner.

By using classification systems which emulate a perceptual evaluation it might be possible to identify new acoustic features or parameterizations which could be used by clinicians as a basis to perform a more objective assessment. The present paper describes an automatic $Roughness(R)$ evaluation system, based in complexity and Modulation Spectra (MS) features.

Complexity measurements have been used in several studies to determine the presence of a pathology related to the phonatory system [4–6] whereas in [7] GRBAS traits are classified using complexity among other parameterizations. Besides, MS has been used in [8, 9] to detect pathological voices. In [10, 11] it was utilized in pathology automatic classification and in [8] to obtain objective parameters to quantify voice quality. In these studies MS is revealed as a source of parameters to characterize pathological voices. On the other hand, works as [12] use acoustic parameters for automatic classification of Breathiness, obtaining a 77% of efficiency whereas [13] uses MFCC in a GRBAS classification system obtaining 65% efficiency. On [14–16], Linear Frequency Spectrum Coefficients (LFSC) are used to classify different traits in order to test the influence of frequency range in GRBAS perceptual and automatic assessments.

In this paper new MS features are proposed. Moreover, dimensionality reduction techniques and Gaussian mixture models (GMM) are employed for taking decisions on the level $(0-3)$ of R trait using the proposed parameters as input.

The paper is organized as follows: Section 2 presents the theoretical background of complexity and modulation spectra features. Section 3 presents the experimental setup and describes the database used in this study. Section 4 presents the obtained results. Finally, section 5 presents the discussion, conclusions and future work.

2 Theoretical Background

2.1 Complexity measures

For extracting complexity measures, it is first necessary to represent the time series in a m-dimensional space, called *phase* or *state space*. In this manner, the dynamical evolution of the system, all its states and its evolution are described. The procedure usually employed is called *embedding*. Through *embedding* it is possible to calculate an *attractor* which is used to obtain the complexity measurements [17].

Some popular features are the Correlation Dimension (CD), the Largest Lyapunov Exponent (LLE) and also some Regularity measurements such as the Approximate Entropy (ApEn), the Sample Entropy (SampEn) and the Fuzzy Entropy (FuzzyEn).

Correlation dimension CD is the autosimilarity of an embedded time series. It is estimated as presented in [17].

Largest Lyapunov Exponent LLE is a measure of the divergence of nearby orbits in phase space, thus measuring the sensitivity to initial conditions of embedded systems. It is estimated as in [17].

Regularity measurements ApEn was proposed in [18], and tries to measure the regularity of a system. Since ApEn is biased due to a phenomena called self-matching, the *Sample Entropy* is proposed in [19]. The *Fuzzy entropy* is a further improvement which changes the measurement function used in ApEn and SampEn by a Fuzzy membership function [20]. All ApEn, SampEn and FuzzyEn rely on the choosing of the tolerance parameter r, which is usually calculated as $r = \alpha\,\mathrm{std}(\cdot)$, where α is varied from within a delimited range and $\mathrm{std}(\cdot)$ is the standard deviation of the time series.

2.2 Modulation Spectra

On this study new MS parameters are proposed to characterize the voice signal. MS provides information about the energy of modulation frequencies that can be found in the carriers of a signal. It is a bidimensional representation where abscissa usually represents modulation frequency and ordinate axis, acoustic frequency. This kind of representation allows observing different voice features simultaneously such as the harmonic nature of the signal and the frequency modulation of fundamental frequency and harmonics. To obtain MS, signal passes through a short-Time Fourier Transform (sTFT) filter bank whose output is used to detect amplitude and envelope. This output is finally analyzed using FFT [21]. To calculate MS, Modulation Toolbox library ver 2.1 has been employed [22].

After obtaining MS it is needed to extract some parameters representative enough to be used in the classification stage. The MS is parameterized using centroids [23] (MSC) and a set of 5 new features: Low Modulation Ratio (LMR) in modulus and Contrast (MSW) and Homogeneity (MSH) in modulus and phase.

Centroids MSC are obtained along the modulation frequency bands. MS is reduced to an entire number of bands usually ranging from 6 to 26. Once the reduced MS is computed, centroids are calculated and normalized taking into account the energy at the voice pitch in acoustic frequency.

Low Modulation Ratio LMR is the ratio between the energy in the first modulation band at pitch frequency, E_0, and global energy in all modulation bands covering at least from 0 to 25 Hz at pitch frequency (acoustic band), E_{25}, as it is shown in equation 1

$$LMR = 10 \cdot \log(\frac{E_{25}}{E_0}) \tag{1}$$

Contrast and Homogeneity Representing MS as two dimensional images it is observed that pathological voices usually seem to have more complex distributions. Images related to normal voices are frequently more homogenous and have less contrast, as can be seen in Fig. 1

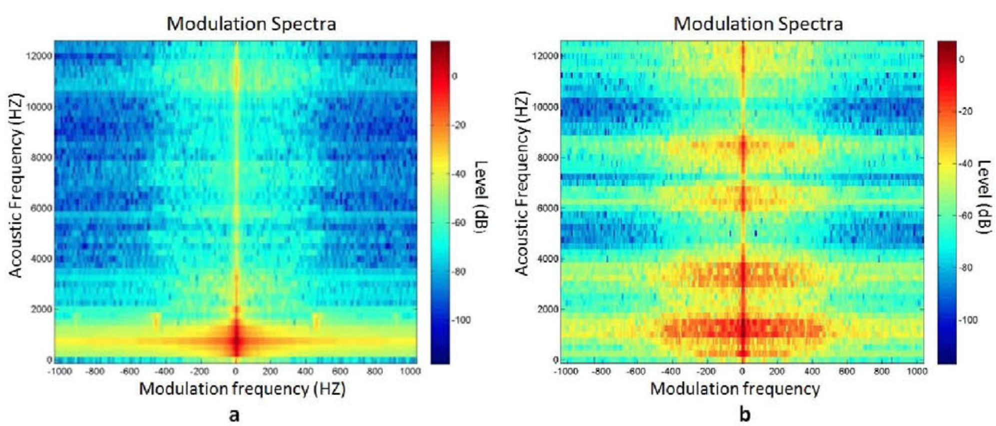

Fig. 1. MS modulus of a normal voice (a) and pathological voice of a patient with gastric reflux, edema of larynx and hyperfunction (b).

Homogeneity is computed using the Bhanu method described by equation 2 as stated in [24].

$$MSH = \sum_{I_m} \sum_{I_a} [f(I_m, I_a) - \overline{f}(I_m, I_a)]^2, \tag{2}$$

being MSH the MS Homogeneity value, $f(I_m, I_a)$ the modulation spectra (modulus or phase) at point (I_m, I_a), and $\overline{f}(I_m, I_a)$ the average value in a 3×3 window centeredred at the same point, representing I_m the frequency modulation bands and I_a the acoustic frequency bands.

Contrast is computed using a variation of the Weber-Fechner contrast relation method described by equation 3 as stated in [24].

$$MSW(I_m, I_a) = \sum_{I'_m} \sum_{I'_a} C_{I'_m, I'_a} \tag{3}$$

where

$$C_{I_m, I_a} = \frac{|f(I_m, I_a) - f(I'_m, I'_a)|}{|f(I_m, I_a) + f(I'_m, I'_a)|} \tag{4}$$

being $f(I_m, I_a)$ MS value (modulus or phase) at point (I_m, I_a) and (I'_m, I'_a) vertical and horizontal adjacent points to (I_m, I_a). The global MSW is considered as the sum of all points in $MSW(I_m, I_a)$.

Regarding MSH and MSW, modulus and phase parameters are used on this study.

3 Experimental Setup

3.1 Database

The original database used for this study contained 226 recordings of maintained vowel /a:/ and the 'Rainbow passage' from the Massachusetts Eye & Ear Infirmary (MEEI) Voice Disorders Database, distributed by Kay Elemetrics [25]. Sample frequency has been adjusted to 25 kHz and quantization to 16 bits when necessary. Duration of the files used for parameterization (only sustained vowel) ranges from 1 to 3 seconds. Level of R trait has been estimated three times by two voice therapists. One of them evaluated the whole database once and the other performed the assessment in two different sessions. Only the 85 files with total agreement among the three assessments were chosen with the aim of using a database with highly consistent labels. This reduced set includes 34 male voices with age ranging from 26 to 58 years with an average of 38 and 51 female with age ranging from 22 to 52 years with an average of 35. Class distribution is shown in figure 2.

3.2 Methodology

The methodology employed in this paper is shown in Fig. 3, while each one of its stages is explained next. Firstly, each signal is framed and windowed using Hamming windows overlapped 50%. The window lengths are varied in the range of

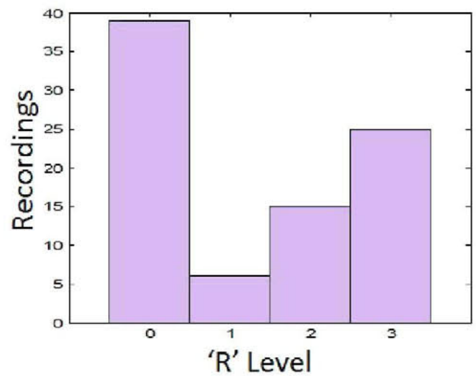

Fig. 2. Class histogram for trait 'R'

40-200 ms in 20 ms steps. Then, in the characterization stage, MS and complexity features are employed. The feature vector extracted from the MS amplitude is composed of the following: MSC, LMR, MSW and MSH. Additionaly, MSW and MSH are computed from the phase. The number of centroids for the MSC feature is varied in the range of $[6, 26]$ with a step size of 2. The complexity set of features is composed by CD, LLE, ApEn, SampEn, and FuzzyEn. The α parameters used for ApEn, SampEn, and FuzzyEn is varied in the following range: $[0.10, 0.35]$ with a step size of 0.05. Following the characterization, a 6-fold cross-validation [26] was used for evaluating results, were two experiments are defined: one without a dimensionality reduction technique, which then feeds the classification stage, and another one using various dimensionality reduction techniques previous classification. In the dimensionality reduction stage PCA [27] and LDA [28] techniques are used, varying the amount of desired features reduction from 25 to 54 %. Regarding PCA and LDA techniques, only the training data set is used to obtain the models which are employed to reshape all the data: training and test data sets. This process is repeated in every iteration of GMM training-test process carried out for validation. The reduction of dimensions is applied for both MS and complexity features separately. Finally in the classification stage, a GMM whose parameters were varied 8 to 128. The assessment of the classifier was performed by means of efficiency and Cohen's Kappa Index [29].

4 Results

Best results can be observed in Table 1. All tests were performed using the described reduced database with and without PCA and LDA techniques. The training set (5 folds from a total of 6) was used to train the GMM models which were validated with the remaining test fold following a 6 fold cross-validation technique.

Fig. 3. Outline of the 'R' automatic detector presented in the paper

Table 1. Best Results. Efficiency ± standard deviation and Kappa Index

Parameters	Efficiency	Kappa Index
Complexity	71 ± 7 %	0.53
Complexity+PCA	67 ± 8 %	0.46
Complexity+LDA	73 ± 8 %	0.58
MS	61 ± 8 %	0.35
MS+PCA	73 ± 15 %	0.56
MS+LDA	**82 ± 7 %**	**0.73**

Best results were obtained using MS in 180 ms frames, 8 centroids, 54 % data reduction through LDA and 14 GMM. Regarding Complexity parameters, best results are obtained with $\alpha = 0.25$, 25 % LDA dimensionality reduction and 16 GMM. All results are expressed in terms of efficiency and Cohen's Kappa Index, the latter expressing the grade of agreement between the labels assigned by the GMM classifier and the perceptual assessment done by therapists.

On Tables 2 and 3 confusion matrices are shown respectively for MS and complexity features. These matrices are the sum of the confusion matrices obtained in each of the six test folds.

Table 2. MS Parameters Confusion Matrix. TR are targets and PR predictions

	PR 0	PR 1	PR 2	PR 3
TR 0	**38**	0	1	0
TR 1	3	**1**	2	0
TR 2	1	0	**10**	4
TR 3	2	0	2	**21**

5 Discussion and Conclusions

On this study an analysis of two different parameterizations applied to human voice to characterize the level of *Roughness*(R) has been performed. Dimension-

Table 3. Complexity Parameters Confusion Matrix. TR are targets and PR predictions

	PR 0	PR 1	PR 2	PR 3
TR 0	**34**	0	0	5
TR 1	2	**1**	2	1
TR 2	2	1	**4**	8
TR 3	2	0	0	**23**

ality reduction methods as LDA and PCA and GMM classification techniques have been used to analyze the capability of both types of parameterizations to characterize voice roughness. Best results are obtained with the proposed new MS parameters and LDA, producing 82 % of efficiency and 0.72 Cohen's Kappa Index. As it can be inferred from Altman interpretation of Cohen's index [30], shown in Table 4, agreement is considered as good. Moreover, most of errors are placed in adjacent classes as it can be deduced from confusion matrices in Tables 2 and 3.

Table 4. Altman interpretation of Cohen's index

Kappa Index	Agreement
$\leq$ 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Medium
0.61 - 0.80	Good
0.81 - 1.00	Excelent

As a starting point, most of the previously exposed tests were performed with the extended database (226 files) using the three available label groups separatedly: one of them generated by one therapist and the other two created by the other therapist in two different sessions. In these cases, on spite of having a higher number of files and a more class-balanced database, results rarely outranged 62 % of efficiency. The details of these tests have not been included in this work for the sake of simplicity and conciseness. This demonstrates that consistency of the database labeling is a key point for future work. New studies should utilize only consistent labels obtained with several therapists in different sessions.

It is difficult to compare these results with other studies such as [12–16] due to, as it is stated in [31], there is not a standard database and in this particular case labeling is different for each work, although results in most of them are under 80% efficiency. The definition of a standard database with a consistent and known labeling would lead to comparable results.

As a conclusion, it might be said that results suggests that the proposed MS parameters could be used as an objective basis to help clinicians to assess

Roughness according the GRBAS scale reducing uncertainty. The use of MS seems to provide better results than complexity. It would be advisable to study the creation of a new parameter from the combination of the proposed ones, being suitable for therapists and physicians. But to obtain highly relevant and representative results, new tests with a larger database should be performed which will allow the use of a a slightly different and more consistent methodology.

6 Acknowledgements

The authors of this paper have developed their work under the grant of the project *TEC2012-38630-C04-01* from the Ministry of Economy and Competitiveness of Spain and *Ayudas para la realización del doctorado (RR01/2011)* from Universidad Politécnica de Madrid, Spain.

References

1. M. Hirano, *Clinical examination of voice.* Springer Verlag, 1981.
2. I. V. Bele, "Reliability in perceptual analysis of voice quality." *Journal of voice : official journal of the Voice Foundation*, vol. 19, no. 4, pp. 555–73, Dec. 2005.
3. W. F. L. De Bodt, M S. *et al.*, "Test-retest study of the grbas scale: influence of experience and professional background on perceptual rating of voice quality." *Journal of voice : official journal of the Voice Foundation*, vol. 11, no. 1, pp. 74–80, 1997.
4. Y. Zhang, J. J. Jiang, L. Biazzo *et al.*, "Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis." *Journal of voice : official journal of the Voice Foundation*, vol. 19, no. 4, pp. 519–28, Dec. 2005.
5. J. J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement." *Journal of voice : official journal of the Voice Foundation*, vol. 20, no. 1, pp. 2–17, Mar. 2006.
6. G.-L. J. I. S.-L. N. Arias-Londono, J. D. *et al.*, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2011.
7. J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón *et al.*, "Automatic GRBAS Assessment Using Complexity Measures and a Multiclass GMM-Based Detector," *Seventh International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2011.
8. M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions On Audio Speech And Language Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.
9. J. D. Arias-Londoño, J. I. Godino-Llorente *et al.*, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices." vol. 36, no. 2, pp. 60–9, Jul. 2011.
10. T. F. Q. Nicolas Malyska, "Automatic dysphonia recognition using biologically inspired amplitude-modulation features," in *Proc. ICASSP*, vol. 1. IEEE, 2005, pp. 873–876.
11. M. Markaki and Y. Stylianou, "Modulation Spectral Features for Objective Voice Quality Assessment: The Breathiness Case," *Sixth International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications.*, 2009.

12. A. Stráník, R. Cmejla, and J. Vokřal, "Acoustic Parameters for Classification of Breathiness in Continuous Speech According to the GRBAS Scale." *Journal of voice : official journal of the Voice Foundation*, vol. 28, no. 5, Sep. 2014.
13. N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz *et al.*, "Automatic assessment of voice quality according to the GRBAS scale," *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 2478–2481, 2006.
14. G. Pouchoulin, C. Fredouille, J. Bonastre *et al.*, "Dysphonic Voices and the 0-3000Hz Frequency Band," *Interspeech 2008. ISCA*, pp. 2214–2217, 2008.
15. G. Pouchoulin, C. Fredouille, J.-F. Bonastre *et al.*, "Characterization of the pathological voices (dysphonia) in the frequency space," *Proceedings of International Congress of Phonetic Sciences (ICPhS)*, pp. 1993–1996, 2007.
16. G. Pouchoulin, C. Fredouille, J. Bonastre, A. Ghio *et al.*, "Frequency Study for the Characterization of the Dysphonic Voices," *Interspeech 2007. ISCA*, pp. 1198–1201, 2007.
17. H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, 2nd ed. Cambridge University Press, 1 2004.
18. S. M. Pincus, "Approximate entropy as a measure of system complexity." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 6, pp. 2297–301, Mar. 1991.
19. J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy." *American journal of physiology. Heart and circulatory physiology*, vol. 278, no. 6, pp. H2039–49, Jun. 2000.
20. W. Chen, Zhuang *et al.*, "Measuring complexity using fuzzyen, apen, and sampen," *Medical Engineering & Physics 31 (2009) 61-68*, 2009.
21. S. Schimmel, L. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," *EEE International Conference in Acoustics, Speech and Signal Processing, 2007. ICASSP*, vol. 4, 2007.
22. L. Atlas, P. Clark, and S. Schimmel, "Modulation Toolbox Version 2.1 for MATLAB," 2010. [Online]. Available: http://isdl.ee.washington.edu/projects/modulationtoolbox/
23. B. Gajic and K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 600–608, Mar. 2006.
24. R. Peters and R. Strickland, "Image complexity metrics for automatic target recognizers," *Automatic Target Recognizer System and Technology Conference*, 1990.
25. "Voice Disorders Database," Lincoln Park, NJ., 1994.
26. B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, 1983.
27. L. Smith, "A tutorial on principal components analysis," *Cornell University, USA*, vol. 51, 2002.
28. R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, pp. 13–16, 1992.
29. J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. XX, no. 1, pp. 37–46, 1960.
30. D. G. Altman, *Practical statistics for medical research.* CRC Press, 1990.
31. N. Saenz-Lechon, J. I. Godino-Llorente, V. Osma-Ruiz *et al.*, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006.